Sampling Approaches for Nutrient Data Bases

S. M. Nusser and A. L. Carriquiry

Department of Statistics and Statistical Laboratory

Iowa State University

Prepared for the

Nutrient Data Laboratory

Beltsville Human Nutrition Research Center

Agricultural Research Service

U.S. Department of Agriculture

Introduction

The USDA Nutrient Data Laboratory (NDL) is developing a plan for updating their nutrient data bases under the auspices of the National Food and Nutrient Analysis Program. One component of this program is to develop a general approach to obtaining samples of foods for nutrient analyses. The resulting data would be used to estimate parameters of nutrient concentration distributions for use in NDL nutrient data bases.

This report describes alternative approaches for sampling food items. We begin by outlining principles of survey sampling and problems associated with sampling foods. We then describe a general framework for sampling foods and discuss adaptations of the approach for classes of food items with varying sampling characteristics. Examples are provided, and possible estimation strategies are outlined.

Priority Food Items and Sampling Resources

The NDL produces two data bases: the Standard Reference Data Base, which contains nutrient values for more than 2500 food items, and the Survey Nutrient Data Base, which contains nutrient values for more than 6600 food items. One thousand of these food items have been identified as the focus of nutrient determinations. Of the 1000 foods, nearly 700 are key foods. The key foods for a nutrient are defined as those food components which together supply up to 80% of the nutrient intake for the population. Key foods as a whole have been found to account for approximately 90% of the nutrient content of the U.S. diet for 30 nutrients of interest (Haytowitz, et al., 1996). The remaining 300 + priority foods include mixed dishes, foods eaten by at-risk populations (e.g., low income, children), ethnic foods, new foods, and ingredients, such as spices.

The NDL has classified the 1000 priority food items into four priority categories (Table 1). Priorities were determined by developing a score for each item that summarizes the total nutrient contribution of that item to the US diet. Those foods that are in the top 25% regarding their contribution to nutrient intake have been classified as belonging to the highest priority, category I. The second, third, and fourth quartiles comprise priority categories II, III, and IV, respectively.

An estimate of the mean and variance for concentrations of each nutrient are desired for each food item. However, because the cost of a complete nutrient profile for one sample of food is approximately $2000, the NDL has calculated that it has resources to run nutrient profiles for, on average, a relatively small number of lab analyses per food. Because of these constraints, the NDL proposes to vary samples sizes across priority categories to achieve different estimation goals for the mean and the variance (Table 1). For priority I items, sample sizes must be large enough to precisely estimate the mean and variance of nutrient content in the food. For priority II, a good estimate of the mean and a less precise estimate of the variance are desired. Even

Table 1. Relationship between Priority Categories, Statistical Precision, and Sample Size

Priority Category	Contribution to nutrient intake	Mean	Variance	Number of samples to be analyzed
I	Top 25%	Robust	Robust	25-50
II	26-50%	Robust	Marginal	12-20
III	51-75%	Robust	Preliminary	3-12
IV	76-100%	Preliminary	Preliminary	3-5

smaller sample sizes are designated for category III foods, for which a good estimate of the mean is desired, but a only rough estimate of the variance is expected. Priority category IV includes those foods in the bottom quartile regarding their contribution to nutrient intake, and rough estimates of the mean and variance will suffice. The precision levels (robust, marginal, preliminary) are relative terms, and are undefined at this time. Some aggregation of the samples for lab analyses is expected, although it is recognized that multiple analytic samples are required to calculate variance estimates.

Fundamentals of probability sampling

Design

Most well-designed scientific studies involve selecting a random sample from a defined target population, observing data on the sample units, and analyzing sample data to make inferences about the target population. Statistical methods used in many areas of scientific inquiry are developed for infinite populations (e.g., experimental design, ANOVAs), and frequently require distributional assumptions to obtain estimators of population parameters. In contrast, survey sampling involves selecting random samples from a *finite* population using a specified sampling design. Distributional assumptions are not required to estimate parameters although models may be used to improve estimation. Cochran (1977), Särndal et al. (1992), and Thompson (1992) are standard references.

In survey samples, the target population can be thought of as a finite list of units or *elements*. Samples are selected from a list referred to as the *sampling frame*, which to the extent possible, includes all elements in the population. When the frame does not fully cover the target population or information used to develop the frame is inaccurate, estimates obtained from the sample may be biased. A frame also includes variables to identify the sample design structures, such as strata or clusters.

A wide variety of sample designs exist for selecting elements from populations. A basic design is simple random sampling, which involves selecting elements from finite populations in a manner such that all samples are equally likely. Stratified random sampling involves dividing the population into mutually exclusive groups of elements called *strata*, and selecting independent random samples within each stratum. Stratification is used to ensure that the sample is spread across the full range of conditions in the population, to obtain adequate sample sizes for populations, to implement varying sample designs across subpopulations, and to address operational constraints.

Another design is two-stage cluster sampling, which involves two sampling steps. During the first stage of sampling, groups of population elements, called *clusters* or *primary sampling units* (PSUs), are selected. Then a subsample of elements is selected from each sample PSU during a second stage of sampling. Two-stage (or multi-stage when more than two stages are used) cluster samples involve different kinds of sampling units at each stage. Cluster sampling is frequently used to improve operational efficiency because clusters can be defined as sets of sample units that are geographically proximal. The statistical efficiency of cluster samples may be reduced when incorporating cost constraints into the design.

Another type of nested design is two- (or multi-) phase sampling. A sample of elements is selected in the first phase, usually to collect data on variables that are relatively simple or cheap to observe. The second phase typically involves selecting a subsample from the first sample, and collecting intensive observations on the second-phase units. The information collected from the two samples are combined during the estimation process. When first- and second-phase variables are highly correlated, this design provides an effective method of compensating for small sample sizes associated with collecting expensive data with limited resources. In addition, multi-phase designs can be used when frames are not available or are too expensive to construct. However, parameter estimation can be quite complicated, especially when separate models are required for each variable.

In practice, these design structures are typically combined to address statistical objectives and operational constraints. For example, a common sample design is stratified multi-stage sampling. The list of clusters is divided into mutually exclusive strata and clusters are sampled within strata. Elements are then selected from sample clusters. The stratification is typically designed to ensure respresentativeness and improve precision, while the clusters are defined to minimize survey costs.

Estimation

Classical survey sampling estimators for population parameters are derived from the properties of the sample design, much like randomization tests are derived in experimental design. These estimators are often referred to as *design-based* estimators, and typically have the form $\sum_{i \in A} w_i y_i$ where y_i is the variable of interest for selected element i, w_i is the *sample weight*, and A is the set of labels for elements contained in the sample. In its simplest form, the sample weight is the inverse of the probability of the element being included in the sample; this probability is called the *inclusion* or *selection probability*. The weight is essentially a measure of how many elements in the population are represented by the sample element. It is possible to construct *self-weighting* designs, where all sample weights are approximately equal, which generally have better statistical properties and can be used to minimize the use of weights in estimation.

Because the sample is selected from a finite population, variance estimators differ from those associated with experimental designs for infinite populations. Standard survey sample variance estimators include a finite population correction term, which is a function of the fraction of population elements sampled. For example, for a simple random sample of size n from a population of size N, an unbiased estimator of the variance of the sample mean is $n^{-1}(1-N^{-1}n)s^2$, where $s^2 = (n-1)^{-1} \sum_{j\in A}(y_j-\bar{y})^2$, $\bar{y}=n^{-1}\sum_{j\in A} y_j$, and A is the set of element indices for the sample. If the population is large, the finite population correction term, $(1-N^{-1}n)$, is approximately equal to one. However, most sample designs have variance estimators that do not reduce to $n^{-1}s^2$. For example, assuming that simple random sampling is used to select elements within each stratum, the estimator for the mean and for the variance of the estimated mean for a stratified random sample are, respectively, $\bar{y}_{st} = N^{-1}\sum_i N_i\bar{y}_i$ and $\hat{V}(\bar{y}_{st}) = N^{-2}\sum_i N_i^2(1-N_i^{-1}n_i)n_i^{-1}s_i^2$, where N_i is the population size in stratum i, $N = \sum_i N_i$ is the total population size, n_i is the sample size in stratum i, $n = \sum_t n_i$ is the total sample size, $\bar{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ is the sample mean for stratum i, and $s_i^2 = (n_i-1)^{-1}\sum_{i=1}^{n_i}(y_{ij}-\bar{y}_i)^2$ is the sample variance for stratum i.

Alternatively, consider a two-stage cluster sample where n clusters are sampled with replacement from a population of N clusters, with selection probability for cluster defined to be

π_j . Suppose a simple random sample of m_j elements is selected from the M_j elements in cluster j. Let $M = \sum_{j=1}^{n} M_j$ be the total number of elements in the population, y_{jl} be the value for element l in cluster j, and $\bar{y}_j = m_j^{-1} \sum_{l=1}^{m_j} y_{jl}$ be the sample mean for cluster j. An unbiased estimate of the mean value of y per element is $\hat{\mu} = M^{-1} \sum_{j=1}^{n} \pi_j^{-1} M_j \bar{y}_j$. The estimated variance of $\hat{\mu}$ contains two variance components that correspond to the between- and within-cluster variation, respectively. Under certain conditions, a two-stage cluster sample can be constructed as a self-weighting design, and the mean and variance estimators reduce to the sample mean and variance, respectively.

Application of Sampling Concepts to Foods

The objective in developing a sampling plan for a specific food is to select units of the food item from the population of units consumed in the U.S. Sample units are sent to a laboratory, and measurements of nutrient content are made on each sample unit or on a pooled set of sample units. These measurements are used to estimate parameters, such as the mean and variance, of the distribution of nutrient concentrations for the specific food item.

To illustrate how the sampling concepts described in the previous section relate to sampling of food items, suppose the objective is to estimate the mean and variance for the amount of calcium per gram of full fat margarine. The population of interest is the collection of boxes (and other units such as tubs) of full fat margarine consumed in the U.S. An element is a single box of margarine. A sample frame is conceptually defined as all full fat margarine boxes available in food outlets throughout the U.S. A sampling unit is a single box of margarine. Given a simple random sample of n boxes, A denotes the set of boxes of full fat margarine included in the sample, n is the number of units in the set A, and y_i represents the calcium content per gram in box i of full fat margarine.

In practice, no list of food items exists that can be used to select a sample of food items that is representative of the population of all units consumed in the U.S. In addition, sampling food items is a challenging task because resources are limited relative to the number and variety of foods and nutrients in the population. Factors such as perishability, seasonality, geographic distribution, and market penetration of different brands have varying effects on sample design and on sample size for each food item. Furthermore, sources and formulations of food items are diverse. For a given food item, the population may include units available in grocery outlets ranging from major chains to small natural food stores, in restaurants ranging from chains to street vendors, or in some combination of grocery and restaurant outlets. There may be only one national brand or producer that supplies the market, or there may be a wide variety of brands, possibly varying a cross region or in relation to local sources. There may also be many formulations of the product, and the specific formulation may be altered over time in response to anticipated market forces. For example, there are over 40 brands of full fat margarine sold in the U.S. These brands are not equally distributed across the U.S., and they vary, for example, in the vegetable oil type, which affects the content of various nutrients. The available formulations of the product can be quite volatile, and consumer preferences for brands and formulations varies geographically.

A stratified multi-stage sampling approach can be used to obtain an operationally feasible and representative sample of foods from the food supply with limited resources. A single sampling plan cannot be developed to accommodate attributes of very different types of food (e.g., ground beef, rice, municipal drinking water). However, a general sampling framework can be outlined which can be adapted to several broad groups of foods that vary in characteristics that affect operational constraints. We now describe the general approach, how to proceed at each stage of sampling, and suggest possible sources of information that may be used to develop an efficient procedure.

General Sampling Approach

Overview

A stratified multi-stage cluster sampling approach is needed to address the complications associated with sampling foods. Sampling stages involve selecting small area segments from the U.S. surface area, then outlets within area segments, and food items within sample outlets.

A representative sample is obtained by stratifying the population at each stage of sampling by factors related to the variation in the population. Factors affecting variation in nutrient content across the U.S. include geography, cultural preferences, outlets where the food is purchased, brand preferences, formulation, and preparation method. First-stage strata for area segments are regions of the U.S. defined by geography and population density. Second-stage sampling units are food outlets, stratified by outlet type. Third-stage elements are food items, stratified by factors such as brand name, formulation, etc.

The clusters in this design are geographic areas (clusters of outlets) and outlets (clusters of units for food items). Clustering is used to achieve operational efficiency. For example, selecting clusters of outlets that are geographically proximal makes it possible to develop a relatively complete list of outlets which covers the breadth of food item sources in the geographic area. Likewise, selecting a handful of outlets for obtaining food items reduces costs for sample pick up.

First-Stage Sample Design

The first stage of sampling involves selecting a stratified area sample. The U.S. surface area is partitioned into area segments, which represent the primary sampling units (PSUs). Area segments should be constructed such that it is operationally feasible to fully enumerate food outlets within the PSU to the extent possible. In rural areas, an area segment may be a county, while in higher population density areas, a smaller area segment may need to be defined.

The list of area segments should be divided into 8-12 strata corresponding to geographic regions and population density. This stratification is designed to obtain physical spread across the country and across population densities. The number of strata will be influenced by operational constraints and population size. The precise definition can be based on broad differences in food purchasing habits using purchase data (e.g., from Nielsen), food consumption behaviors using food consumption surveys (e.g., from the USDA 1994-1996 CSFII), and demographic information (e.g., from U.S. Bureau of the Census). At least one stratum should consist of low population density area segments to ensure representativeness for food items, such as municipal drinking water or locally produced products, where differences in nutrient content might be expected to vary by population density. In addition, it may be desirable to define strata to include large cities that should be included in the sample with certainty or with a high probability. For example, one stratification scheme might include three high population density strata for large cities in the western, central, and eastern parts of the U.S.; two low density strata for the east and west; and seven additional strata that are defined by geographically proximal area segments. Given the second- and third-stage sample designs proposed in later sections, it may be useful to define strata such that the strata have roughly equal population sizes in order to avoid creating highly variable sample weights.

Area segments should be selected with probability proportional to population within the PSU as part of the strategy to develop an approximately self-weighting design. The frame used to select area segments should contain the list of PSUs, the stratum to which each PSU is assigned, and PSU inclusion probabilities proportional to the population size in each PSU (derived from Census data). The frame in each segment would remain in use for several years and should be updated as needed.

Regarding sample sizes, a limited number of area segments per stratum is suggested so that considerable effort can be devoted to developing a good outlet frame within the PSU as the basis for creating sampling plans for numerous priority food items. To ensure that the sample is dispersed across geographic regions and population densities, it may be useful to apply a one-per-stratum design for PSUs, in which only one area segment is selected per stratum. Thus, 8-12 PSUs would be selected in the first stage of sampling. For future reference, let M_i denote the total number of PSUs in stratum i, and m_i denote the number of PSUs selected within stratum i.

Second-Stage Sample Design

In the second stage of sampling, a sample of food outlets is selected from each sample PSU. The outlet represents the secondary sampling unit (SSU). A comprehensive list of all outlets in each PSU must be assembled in preparation for sampling. Although this will require considerable effort, the outlet frame within each PSU is the cornerstone to developing customized sample designs for individual foods. Its repeated use justifies spending time on constructing a good frame. Note that one frame is needed for each of the 8 - 12 PSUs in the first-stage sample.

Assembling the frame involves developing a list of outlets, classifying the outlets for stratification, and determining a "size" measure related to propensity to consume food from an outlet for use in calculating outlet inclusion probabilities. Listings of outlets are available from various sources of market information. However, these lists tend to be restricted to metropolitan areas and may include only outlets that sell large volumes of the food item. Alternatively, a relatively current and complete outlet list may be obtained from yellow pages that cover the areas located within the PSU. Yellow pages may contain outlets that are located outside of the boundaries of the PSU, and these outlets should be removed from the frame. Yellow pages may also be missing some outlets, though it is expected that this type of frame error will generally be restricted to smaller establishments. The market survey listings may provide information that can be used to check and complete the outlet frame.

In constructing a sampling frame for food outlets, all outlets should be classified by type, (e.g., major supermarket chains, hamburger restaurants with fast service, institutional cafeterias, etc.). This information is needed to stratify the outlets in different ways depending on the food item to be sampled. A fine classification should be developed for the frame to provide flexibility in creating second-stage samples from each PSU. One might consider 10-20 strata, consisting of, for example major supermarket chains, convenience and small grocery outlets, family-style restaurants, institutional cafeterias, fast-food restaurants (pizza, hamburger, Chinese, Mexican, fish, Italian), bakeries, ice-cream parlors, among others.

It is also desirable to obtain sales volume or a surrogate measure of market share for each outlet for inclusion in the frame. This may be difficult and possibly impractical because information on sales volume is spread across numerous sources and sometimes not readily available. The most comprehensive source of information is the Nielsen data base, which covers much of the grocery store market, but does not contain a complete enumeration of the list of outlets because it is derived from grocery outlets that capture electronic sales data. Restaurant sales data are not uniformly available from a primary source, nor does it necessarily make sense to combine restaurant sales with grocery sales figures for determining inclusion possibilities, although stratification by outlet type may provide a solution to this problem.

In this report, we use N_{ijk} to denote the total number of food outlets in the second-stage stratum k, PSU j, and first-stage stratum i. If the number of food outlets (N_{ijk}) were known for each PSU in the U.S., then $\sum_{ijk} N_{ijk} = N$ would be the total number of food outlets in the U.S. The number of sample food outlets in first-stage stratum i, PSU j and second-stage stratum k is denoted by n_{ijk}.

Up to this point in the sampling process, methods for selecting PSUs (area segments) and constructing the full SSU (outlet) list frame are common for all food items. However, the sample size n_{ijk} and the sample design for selecting the sample outlets from each PSU will vary depending upon the characteristics of individual food items. The set of outlets included in the frame and the stratification strategies to ensure representativeness of the sample will depend on the food item. Not all strata will be used to create an outlet sample for a particular food. For example, no bakeries are needed to sample pizza. Additionally, pizza is purchased from a wide variety of outlets, and in many cases, strata will be grouped prior to selecting the sample. It may be desirable to group outlets into a relatively small number of strata, with the objectives of encouraging representativeness in the sample and of sampling two or more outlets in most of the strata for the purposes of variance estimation. For example, for margarine, outlet strata within a PSU may be defined by large grocery store chain #1, large grocery store chain #2, and all other grocery outlets. If market share information is available, then the outlet sample size should be allocated across strata proportional to stratum market share, and outlets should be selected with inclusion probability π_{ijkl} proportional to the market share of outlet l in second-stage stratum *k*, PSU j and first-stage stratum i. Note that all outlets that serve as a source for the food item should be assigned a positive inclusion probability, whereas ineligible outlets should be assigned inclusion probabilities of $\pi_{ijkl} = 0$.

This approach to stratification and to defining inclusion probabilities ensures that non-traditional outlets such as small grocery stores have a chance to be included in the sample. It can also be used to create an approximately self-weighting design for food items. Further investigation is needed to determine whether surrogates can be developed for outlet market share information when it is unavailable. For example, it may be possible to develop an approximate measure for market share based other factors such as physical size of the outlet or seating capacity. Alternatively, assigning outlets to general levels of market share may be more operationally feasible.

The number of sample outlets in each PSU, $\sum_{k} n_{ijk}$, will depend on the characteristics of the food and its priority category (Table 1). However, it is expected that a relatively small number of outlets (perhaps four to eight) would be selected per food in each PSU. It may be desirable and more operationally efficient to construct a scheme for selecting outlets that incorporates sample design considerations for multiple foods simultaneously. Under this scenario, a larger number of outlets (e.g., 10-25) would be selected per PSU depending on the number and type of target foods in the set for which the outlet sample is designed.

Third-Stage Sample Design

In the third stage of sampling, food units are selected for a specific food item from the sample outlets. If possible, information should be obtained on product availability from outlets sampled in a PSU. These data would be used to identify outlets that serve as sources for food items within the PSU and would provide information to define third-stage strata for selecting individual food items from sample outlets. Construction of third-stage strata requires information on factors such as brands, production sources, and formulation for the given food item, and thus this step in the design is likely to involve some investigation and approximation to complete missing information. An alternative approach to gathering product lists and menus from outlets is to make assumptions about the availability of products in outlets on the basis of outlet type, and to increase sample sizes for food items in outlets to accommodate occasional nonresponse (lack of product) in outlets. At some point in the process, it will be desirable to combine lists from outlets across the country to create an integrated stratification and sample size allocation strategy across the U.S.

Stratification will vary with the food item, and a simple stratification with a small number of strata should be adopted to minimize effort. For example, for a food with two dominant brands, two strata would be defined by the top two brands, and a third stratum would consist of the remaining products for the food. For foods that have a sizable control brand market, all control brands could be considered as a single stratum and remaining brands could be combined to form a second stratum. If possible, the sample size should be allocated across third-stage strata proportional to market share or some surrogate measure of market share.

Once the food item design has been constructed, the final step in the sampling process is to select individual food items from each third-stage stratum. In this design, individual food items are the tertiary sampling units (TSUs) and represent the population elements. Let R_{ijkls} denote the total number of TSUs in the third-stage stratum s, SSU (outlet) l, second-stage stratum k, PSU (area segment) j, and first-stage stratum i, and let r_{ijkls} denote the number of TSUs included in the sample. For a food item, $r = \sum_{ijkls} r_{rjkls}$ is the sample size (number of units of food) collected across the U.S. If the stratum sample size allocation and sample unit inclusion probabilities suggested previously for each stage of sampling have been used, then selecting an equal number of samples within each outlet for a particular food item will result in an approximately self-weighting design (ignoring compositing of food samples).

Measurement of Nutrient Concentrations

Under the proposed design, more food units will be collected than can be analyzed. For most, if not all priority foods, at least some of the units will need to be composited for analysis. If individual units are sent to the laboratory for analysis, a value y_{ijklst}, for $t = 1, \ldots, r_{ijkls}$ samples, will be obtained representing the content of a nutrient per gram of food item t. If compositing of sample food units occurs, then the number of analytical samples, say $\tilde{r}$, sent to lab for analysis is less than the total number of elements, r, in the sample.

Whether from r individual samples, from $\tilde{r}$ composited samples, or from a combination of composite and individual samples, the objective is to obtain an estimate of μ_y, the average content of a nutrient in the food item, and of σ_y^2 , the variance of the nutrient content per gram of the food item.

Compositing Samples

If an individual food item in the sample is sent to a laboratory for nutrient content analysis, then the nutrient concentration is defined to be y_{ijklst} , obtained from food item t bought in outlet l in second-stage stratum k within area segment j in first-stage stratum i . Because nutrient profiling can be quite costly, samples are often combined, or *composited*, to reduce the number of analyses performed. A composite sample is constructed by thoroughly mixing together several individual sample units. Compositing can be effective when the cost of individual lab analyses is high, the mixing into a composite is thorough, and direct information on between-unit variance or extreme values of a variable in the population is not needed.

For nutrient data base parameters, variance estimates are required. One approach is to form one or more composite samples using a subset of the sampled individual food units, and then analyze the remaining individual food units separately. If the number of individual samples used to create the composite samples is known, both the composite and individual sample values can be used to estimate the variance between food units. This method requires weights to be used when estimating means and variances. In addition, small sample sizes of individual measurements may not produce precise estimates of the variance. An alternative approach is to develop several composite samples from the individual food items, leaving no individual samples to be analyzed. A normal distribution can be postulated for the nutrient content of individual samples, and restricted maximum likelihood estimation (REML) can be used to estimate the between-unit variance. It is unlikely that a normality assumption will be suitable for all nutrient concentrations, but this approach may be useful for lower priority foods. Further investigation is needed to determine an appropriate approach for compositing.

Examples

The number n_{ijk} of outlets to be selected within each stratum, the number of third-stage strata with outlets for each food item, and the number of food units to be purchased from the sample outlet depends greatly on the food item. Both the intrinsic characteristics of the food item and its priority category as determined by NDL need to be taken into account.

As an example, consider two food items, full fat margarine (over 80% fat) and thin-crust cheese pizza. Both food items have distinct Nutrient Data Base numbers, and have been classified as priority I foods. Because of the characteristics of these food items, determining the number r_{ijkls} of food units to be purchased from each selected outlet is complex. For thin-crust cheese pizza, selection of outlets itself can be difficult.

Consider first the case of full fat margarine. Margarine is purchased only in grocery outlets. Thus, appropriate third-stage strata in an area segment might include (1) major supermarket chains, and (2) small groceries and convenience stores. Suppose that in the area segment, outlets in stratum 1 account for 80% of total sales of groceries, and outlets in stratum 2 account for the remaining 20%. Here, volume sold is recorded in dollar amounts. This information can be used to decide the number n_{ij1} and n_{ij2} of outlets in each stratum to be selected. For example, for full fat margarine, it would be reasonable to set $n_{ij1} = 4$ and $n_{ij2} = 1$, so that outlet selection from each stratum is done with probability proportional to the stratum's market share. Within the first stratum, the four major supermarkets are selected from among the N_{ij} elements in the stratum with probability proportional to market share. An ordered systematic selection scheme could be used to encourage selection of outlets from different chains in stratum 1.

At this stage, we have made use of market share information for each outlet in the frame. If these data are not available for all frame elements, selection of outlets could result in less precise estimates of nutrient contents. Notice that even though this sampling plan is developed for full fat margarine, market share data has been used for all products sold in the outlets. This is done for simplicity and because it is impractical to obtain each outlet's market share data for each food item of interest.

Once outlets from each stratum have been selected, a product list is obtained from the outlets if possible. This list is used to classify individual units of full fat margarine into third-stage strata prior to selection. For margarine, brand name is the appropriate stratification. According to Nielsen information, about 10 brands (including control brands) dominate the market. This number is large enough to consider combining brands to form strata and/or to consider purchasing a relatively high value r_{ijkl} of food units in each outlet. For margarine, where approximately 50 analytic samples are desired, compositing of some of the individual elements will be required. Suppose 12 PSUs are selected, along with 5 outlets per PSU. If in each outlet we define four strata, and select two products per strata, then the plan results in 12 PSUs × 5 outlets per PSU × 8 samples per outlet, or $r = 480$ units of full fat margarine to be bought nationwide. This number clearly exceeds the maximum number of samples that can be sent to the laboratory for profiling. To obtain a relatively precise estimate of the mean and variance, one strategy would be to create 28 composite samples (obtained by combining 12 individual units of margarine, one from each PSU). Then data from the remaining 24 individual samples would be combined with data from the composite samples for variance estimation. A total of $\widetilde{r} = 52$ analytic samples of margarine would be sent to the lab.

Consider now the case of thin-crust cheese pizza. Stratifying outlets for pizza is more complex than for margarine, since pizza is sold in a variety of forms (frozen, fresh) from a variety of outlets (major grocery stores, other grocery stores, institutional cafeterias, Italian restaurants, pizza restaurants with fast service). Typically, frozen pizza is not sold in restaurants, but this is not always the case (e.g., the Pizzeria Uno chain). Because of this added complexity, thin-crust cheese pizza requires a more elaborate stratification at the second and third stages than margarine.

Within a PSU, stratification might consist of dividing outlets into (1) major grocery chains, (2) major pizza restaurant chains, (3) local pizza restaurants, institutional cafeterias, small grocery and convenience stores, and any other establishment. As before, stratum sample sizes are allocated proportional to volume and outlets within a stratum are selected with probability proportional to the outlet's market share. In stratum 3, where market share is unlikely to be available, one or two outlets might be selected using simple random or systematic sampling. Since three outlet strata are defined within each area segment, proportional allocation of sample sizes to strata (proportional to individual stratum market share) might again result in a relatively large number of outlets to be visited in an area segment. For example, suppose that approximately 30% of all thin-crust cheese pizza is sold by outlets in stratum (1) in a segment area, and that the percentages of strata (2) and (3) are respectively 60% and 10%. Proportional allocation would require that we select, for example, 3, 6, and 1 outlets from strata (1), (2), and (3), respectively. Thus, 10 outlets are selected from each of the 12 area segments, for a total of 120 outlets nationwide. If this were too large a second-stage sample size, smaller samples could be selected within each stratum (e.g., $n_{ijk} = 2, 3, 1$ for $k = 1, 2, 3$, respectively) or strata could be further collapsed.

Concluding Remarks

Laboratory analyses for nutrient profiles are extremely expensive. Relative to lab analyses, sampling is a relatively small component of the cost structure. To maximize the value of the lab analyses, it is of interest to obtain samples of food items that have good properties for estimation and that can be selected in an efficient manner. Our objective was to outline an approach for selecting samples of food items that will be statistically defensible and operationally feasible. The proposed plan represents an initial step in attaining that goal. However, some components of the approach require further investigation to define specific procedures that can be efficiently implemented for hundreds of foods and that provide a reasonable approximation to standard statistical design principles.

The materials and considerations needed to establish the specific parameters of the first-stage area sample design are readily available. Analysis of Census Bureau population data is needed to select the number of strata and to properly define strata. In addition, this information is needed to choose a practical size for the area segment.

Some investigation into available outlet information will be required for area segments. While this work can proceed prior to selecting area segments, it may save effort in the long-run to have selected a PSU sample prior to beginning the work. The first area to consider is how to construct a relatively complete list frame of outlets using yellow pages or other materials. The second objective will be to identify variables that are correlated with market share and readily available for all outlets so that allocation strategies and inclusion probabilities can be determined. The variables may be relatively crude classifications or specific dollar volumes.

The third-stage sample design also requires investigation into materials available for deciding on which products are available and where they can be purchased. Once again, it is important to define an operationally efficient mechanism for selecting individual food items of varying kinds. This stage will likely require more approximation to standard survey methods than the previous stages. It will be useful to identify several different kinds of foods to use in a pilot study of the procedures.

Approaches to compositing also need further study. More than one approach may be needed depending on the precision requirements for variance estimates.

Finally, we have described procedures in a sequential fashion so that steps at each stage could be understood. In practice, it may be useful to create designs with several foods in mind. Thus, a method of outlet selection that supports this kind of approach should be developed.

Literature

Cochran, W. G. 1997. *Sampling Techniques*. Wiley, New York. 428 pp.

Haytowitz, D. B., P. R. Pehrsson, J. Smith, S. E. Gebhardt, R. H. Matthews, B. A. Anderson. 1996. Key foods: setting priorities for nutrient analysis. *Journal of Food Composition and Analysis*, 9:331-364.

Särndal , L. E., Swensson, B., and Wretman, J. 1992. *Model-Assisted Survey Sampling*. Springer-Verlag, New York. 694 pp.

Thompson, S. K. 1992. *Sampling*. Wiley, New York. 343 pp.